J M S K
Jurnal Matematika, Statistika & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 14, No. 1, 54-60, Juli 2017

p-ISSN: 1858-1382
e-ISSN: 2614-8811

Spline Polynomial Truncated dalam Regresi Nonparametrik

Anna Islamiyati*

Abstrak

Model pendekatan regresi nonparametrik yang diberikan oleh persamaan $y_i = f(t_i) + \varepsilon_i,\ \varepsilon_i :\ N\left(0, \sigma^2\right), i = 1, 2, \text{K}, n$ dimana y_i adalah variabel respon sedangkan fungsi $f(t_i)$ merupakan kurva regresi yang bentuknya disumsikan tidak diketahui dan termuat dalam ruang Sobolev $W_2^m[a,b]$. Dalam makalah ini diberikan pendekatan spline *polynomial truncated* yaitu spline dengan orde *p* dengan knots k_1, k_2, K, k_m. Kemampuan spline *polynomial truncated* dalam mengestimasi perilaku data yang cenderung berbeda pada interval yang berlainan ditunjukkan oleh fungsi *truncated* yang melekat pada estimator yang ditunjukkan pada

$$y_i = f(t_i) + \varepsilon_i = \sum_{j=0}^{p} \beta_j t_i^j + \sum_{j=1}^{m} \beta_{j+p} (t_i - k_j)_+^p + \varepsilon_i.$$

Pemilihan titik knots optimal berdasarkan metode *Gross Cross Validation* (GCV). Sebuah aplikasi spline *polynomial truncated* pada data jumlah penjualan roti CV DEDE Makassar. Berdasarkan analisis diperoleh spline orde 3 (spline kubik) dengan empat knots optimal yaitu pada bulan ke-3, 6, 8, dan 11 dengan persamaan estimasi:

$$\hat{y}(t) = 610,202 + 36,411t - 83,508t^2 + 17,486t^3 - 77,378(t-4)_+^3 + 121,429(t-6)_+^3 + -92,575(t-8)_+^3 + 856,159(t-11)_+^3.$$

Kata Kunci: GCV, knots, regresi nonparametrik, spline polynomial truncated.

1. Pendahuluan

Analisis regresi telah menjadi salah satu metode statistika yang sangat berperan dalam perkembangan ilmu statistika khususnya dalam melihat pola hubungan pasangan data antara variabel prediktor dengan variabel respon. Dalam analisis regresi terdapat dua pendekatan yang biasa digunakan untuk mengestimasi kurva regresi, yaitu pendekatan regresi parametrik dan regresi nonparametrik. Pendekatan regresi parametrik digunakan jika bentuk kurva regresi diketahui. Sedangkan pendekatan regresi nonparametrik digunakan apabila informasi mengenai bentuk dan pola hubungan antara variabel prediktor dengan variabel respon tidak diketahui (Budiantara, 2001).

* *Prodi Statistika, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin*

Jika diberikan pasangan data $(t_i, y_i), i = 1, 2, \ldots, n$ dan hubungan antara variabel respon dengan variabel prediktor tidak diketahui bentuknya, maka dapat digunakan pendekatan regresi nonparametrik, dengan model sebagai berikut:

$$y_i = f(t_i) + \varepsilon_i, \ \varepsilon_i : \ N\left(0, \sigma^2\right), i = 1, 2, \text{K}\ , n \tag{1}$$

dimana y_i adalah variabel respon sedangkan fungsi $f(t_i)$ merupakan kurva regresi dengan t_i sebagai variabel prediktor dan ε_i adalah residual berdistribusi normal independen dengan mean nol dan varians σ^2 (Wahba, 1990).

Model regresi nonparametrik memuat asumsi bahwa bentuk kurva regresi *f* tidak diketahui, tetapi hanya diasumsikan termuat dalam suatu ruang fungsi tertentu. Beberapa model pendekatan kurva regresi nonparametrik diantaranya adalah spline. Spline merupakan fungsi yang diperoleh dengan meminimumkan *penalized least square* (PLS) yaitu kriteria estimasi yang menggabungkan *goodness of fit* dengan fungsi penalti (Wahba, 1990). Misalkan diberikan model (1), maka estimator kurva regresi *f* diperoleh dengan meminimumkan PLS:

$$\left\{ n^{-1} \sum_{i=1}^{n} \left(y_i - f\left(t_i\right) \right)^2 + \lambda \int_a^b \left[f^{(m)}\left(t\right) \right]^2 dt \right\} \tag{2}$$

untuk setiap *f* di dalam ruang *Sobolev* $W_2^m [a,b] = \left\{ f \left| \int_a^b \left[f^{(m)}\left(t\right) \right]^2 dt < \infty \right. \right\}$.

Dalam makalah ini, diuraikan model spline, khususnya spline polynomial truncuted yang memiliki fleksibilitas yang tinggi dan kemudahan dengan metode spline akibat pendekatan polynomial yang digunakan sebagai pendekatan fungsi dan interpolasi, sehingga memiliki perhitungan matematika yang sederhana (Weinert dan Kailath, 1974). Bagian 2 disajikan tentang konsep spline *polynomial truncated* dengan penggunaan titik knots dalam perolehan model spline optimal. Bagian 3 disajikan contoh aplikasi dari spline *polynomial truncated* pada penjualan roti pada perusahaan roti DEDE di Makassar dalam setiap bulan penjualan.

2. Spline Polynomial Truncated

Spline adalah potongan polinomial order *p* dengan titik bersama dari potongan-potongan tersebut disebut dengan knots. Titik knots merupakan perpaduan dua kurva yang menunjukkan pola perubahan perilaku kurva pada selang yang berbeda (Hardle, 1990). Spline orde *p* dengan knots pada $k_1, k_2, \text{K}\ , k_m$ diberikan dalam fungsi *f* dengan bentuk:

$$f(t) = \sum_{j=0}^{p} \beta_j t^j + \sum_{j=1}^{m} \beta_{j+p} (t - k_j)_+^p, \ , \tag{3}$$

dengan $\beta_0, \beta_1, \text{K}\ , \beta_{p+j}$ adalah parameter dan $(t-k_j)_+^p = \begin{cases} (t-k_j)^p & ,(t-k_j) \geq 0 \\ 0 & ,(t-k_j) < 0 \end{cases}$

(Eubank, 1988).

Fungsi (3) merupakan fungsi *spline polynomial truncated* yang mempunyai sifat sebagai berikut:

1. Fungsi *f* merupakan potongan polinomial derajat *p* pada setiap subinterval $[k_j, k_{j+1}]$.

2. Fungsi *f* mempunyai turunan kontinu tingkat (*p*-1).
3. f^p merupakan fungsi tangga dengan titik-titik lompatan k_1, k_2, K, k_m.

Apabila diberikan suatu basis untuk spline (Budiantara, 2001) berbentuk $\{1, t, ..., t^p, (t-k_1)_+^p, ..., (t-k_m)_+^p\}$, maka model regresi *spline polynomial truncated* dapat ditulis menjadi:

$$y_i = f(t_i) + \varepsilon_i = \sum_{j=0}^{p} \beta_j t_i^j + \sum_{j=1}^{m} \beta_{j+p} (t_i - k_j)_+^p + \varepsilon_i \qquad (4)$$

dimana *p* adalah derajat *polynomial* dan *m* adalah banyaknya titik knots pada fungsi *truncated* serta ε_i adalah residual berdistribusi normal independen dengan mean nol dan varians σ^2. Akibatnya y_i akan berdistribusi normal dengan mean $f(t_i)$ dan varians σ^2

Model regresi *spline polynomial truncated* pada (4) dapat disajikan dalam bentuk matriks sebagai berikut :

$$\underset{\sim}{y} = \Pi\left[k_1, ..., k_m\right]\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \qquad (5)$$

dimana:

$\underset{\sim}{y} = \left[y_1, y_2, ..., y_n\right]^T$ merupakan vektor variabel respon,

$\underset{\sim}{\beta} = \left[\beta_0, \beta_1, ..., \beta_p, \beta_{p+1}, ..., \beta_{p+m}\right]^T$ vektor parameter yang tidak diketahui,

$\Pi[k_1, ..., k_m]$ matriks berukuran $n \times (p+m+1)$ yang tergantung pada titik knots,

$$\Pi[k_1, ..., k_m] = \begin{bmatrix} 1 & t_1 & \cdots & t_1^p & (t_1-k_1)_+^p & \cdots & (t_1-k_m)_+^p \\ 1 & t_2 & \cdots & t_2^p & (t_2-k_1)_+^p & \cdots & (t_2-k_m)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & \cdots & t_n^p & (t_n-k_1)_+^p & \cdots & (t_n-k_m)_+^p \end{bmatrix}, \text{ dan}$$

$\underset{\sim}{\varepsilon} = \left[\varepsilon_1, \varepsilon_2, ..., \varepsilon_n\right]^T$ vektor residual.

Estimasi parameter model regresi *spline polynomial truncated* yang diperoleh melalui metode *Least Square* melalui optimasi dari:

$$\underset{\underset{\sim}{\gamma} \in R^{p+m+1}}{Min} \left(\underset{\sim}{\varepsilon}^T \underset{\sim}{\varepsilon}\right) = \underset{\underset{\sim}{\gamma} \in R^{p+m+1}}{Min} \left(\underset{\sim}{y} - \Pi[k_1, ..., k_m]\underset{\sim}{\gamma}\right)^T \left(\underset{\sim}{y} - \Pi[k_1, ..., k_m]\underset{\sim}{\gamma}\right) \qquad (6)$$

Diketahui $\Pi[k_1, ..., k_m]$ matriks dengan rank penuh, maka estimasi parameter model ($\hat{\underset{\sim}{\beta}}$) diperoleh:

$$\hat{\underset{\sim}{\beta}} = \left(\Pi^T\left[k_1, ..., k_m\right]\Pi\left[k_1, ..., k_m\right]\right)^{-1}\left(\Pi^T\left[k_1, ..., k_m\right]\right)\underset{\sim}{y} \qquad (7)$$

Estimator *spline* dalam regresi nonparametrik memiliki sifat fleksibilitas yang tinggi dan kemampuan mengestimasi perilaku data yang cenderung berbeda pada interval yang berlainan (Eubank, 1988 dan Budiantara, 2006). Kemampuan ini ditunjukkan oleh fungsi *truncated* (potongan) yang melekat pada estimator yang ditunjukkan pada (4). Selanjutnya estimasi model diperoleh:

$$\underset{\sim}{\hat{y}} = \Pi[k_1,\ldots,k_m]\left(\Pi^T[k_1,\ldots,k_m]\Pi[k_1,\ldots,k_m]\right)^{-1}\left(\Pi^T[k_1,\ldots,k_m]\right)\underset{\sim}{y} \quad = A\left[k_1,\ldots,k_m\right]\underset{\sim}{y} \quad (8)$$

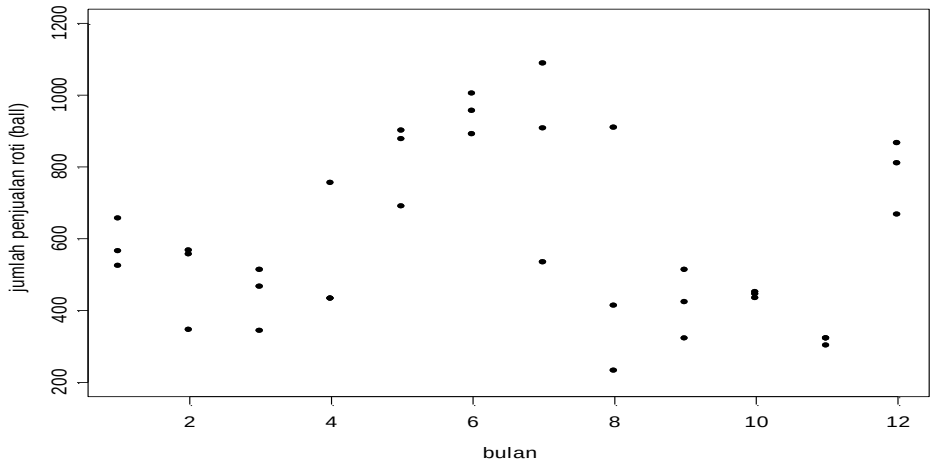
Mengestimasi kurva regresi nonparametrik dengan pendekatan *spline polynomial truncated*, berarti mencari model *spline polynomial truncated* optimal berdasarkan titik knots optimal yaitu berapa banyak titik knots dan dimana letak titik-titik knots tersebut.

Dalam pendekatan model *spline* titik knots harus dipilih dengan berbagai metode seperti *Cross Validation* (*CV*) (Craven dan Wahba, 1979), *Generalized Maximum Likelihood* (*GML*) (Wang, 1998), *Generalized Cross Validation* (*GCV*) dan lain sebagainya. Pemilihan titik-titik knots yang optimal menggunakan metode GCV dinyatakan oleh (Wang, 1998) sebagai berikut :

$$GCV(k_1,\ldots,k_m) = \frac{n^{-1}y^T\left[I - A[k_1,\ldots,k_m]\right]^T\left[I - A[k_1,\ldots,k_m]\right]y}{(n^{-1}trace\left[I - A[k_1,\ldots,k_m]\right])^2} \quad (9)$$

3. Aplikasi

Data jumlah penjualan roti DEDE produksi CV. DEDE Makassar berdasarkan bulan penjualan selama 3 tahun (2004 – 2006). Data diperoleh berdasarkan pencatatan pihak perusahaan, dengan melibatkan beberapa faktor, namun dalam tulisan ini akan ditekankan pada jumlah penjualan setiap bulannya. Plot jumlah penjualan roti dalam setiap bulan diberikan pada Gambar 1.



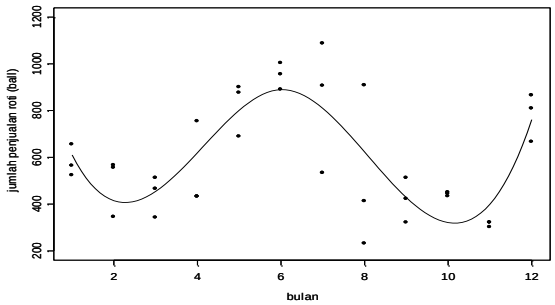
Gambar 1. Plot Data Jumlah Penjualan Berdasarkan Bulan.

Berdasarkan Gambar 1, terlihat bahwa penjualan roti mengalami peningkatan pada awal tahun sampai bulan Juni, kemudian menurun pada bulan Juli sampai November, dan kembali meningkat pada bulan Desember. Ini menunjukkan bahwa plot data tidak mengikuti sebuah kurva parametrik, sehingga untuk mengestimasi jumlah penjualan akan digunakan pendekatan regresi nonparametrik.

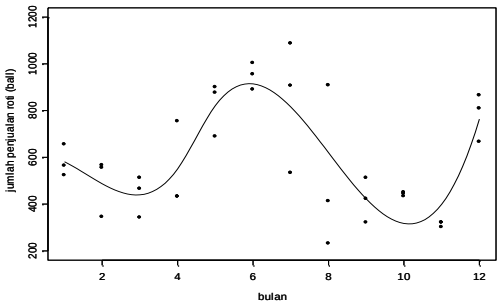
Setelah data dianalisis, model optimal diperoleh pada spline polynomial orde tiga yaitu spline kubik, seperti yang terlihat pada Tabel 1. Tabel 1 menyajikan nilai GCV dari setiap titik knots yang terpilih pada spline kubik. Untuk 1 knots, GCV terkecil yaitu 28.401,1 bersesuaian dengan titik knots k = bulan ke-6. Untuk 2 knots, nilai GCV terkecil yaitu 28.416,2 bersesuaian dengan titik knots k_1 = bulan ke-4 dan k_2 = bulan ke-5. Untuk 3 knots, nilai GCV terkecil yaitu 30.044,3 bersesuaian dengan titik knots k_1 = bulan ke-4, k_2 = bulan ke-5, dan k_3 = bulan ke-8. Sedangkan untuk 4 knots, nilai GCV terkecil yaitu 28.126,5 bersesuain dengan titik knots k_1 = bulan ke-4, k_2 = bulan ke-6, k_3 = bulan ke-8 dan k_4 = bulan ke-11. GCV terkecil terjadi pada penggunaan 4 titik knots yaitu 28.126,5 yang berarti ada empat titik knots optimal dalam spline kubik yaitu pada bulan ke – 4, 6, 8, dan 11 untuk jumlah penjualan roti DEDE di Makassar.

Tabel 1. Nilai GCV pada Berbagai Titik Knots dari Model Spline Kubik.

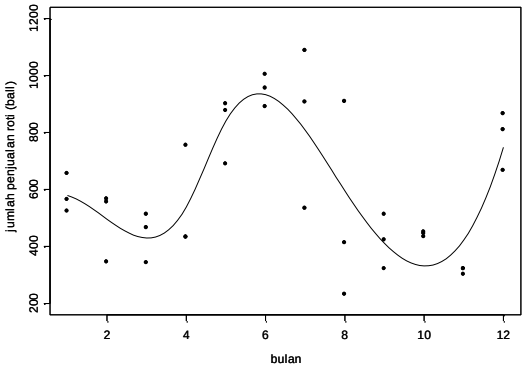
Spline Kubik 1 Knots		**Spline Kubik 2 Knots**			**Spline Kubik 3 Knots**				**Spline Kubik 4 Knots**				
K	GCV	k_1	k_2	GCV	k_1	k_2	k_3	GCV	k_1	k_2	k_3	k_4	GCV
3	44.162,8	3	5	28.786,1	3	4	5	30.406,9	3	6	8	10	28.890,4
4	36.773,9	**4**	**5**	**28.416,2**	**4**	**5**	**8**	**30.044,3**	3	6	8	11	28.568,6
5	30.705,5	3	6	29.381,1	3	6	8	30.190,5	**4**	**6**	**8**	**11**	**28.126,5**
6	**28.401,4**	4	7	31.978,8	4	5	6	30.262,0	5	6	8	10	28.665,5



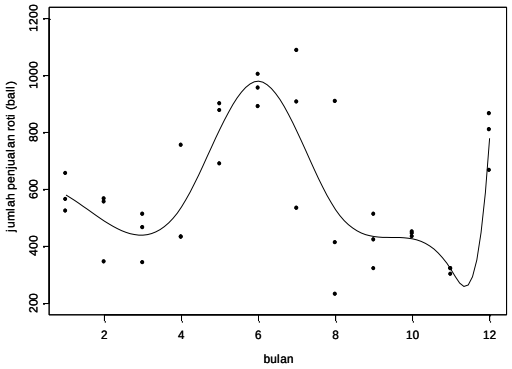
Gambar 2. Spline Kubik dengan 1 Knots (k = bulan ke-6)



Gambar 3. Spline Kubik dengan 2 Knots (k_1 = bulan ke-4 dan k_2 = bulan ke-5)



Gambar 4. Spline Kubik dengan 3 Knots (k_1=bulan



Gambar 5. Spline Kubik dengan 4 Knots

ke-4, k_2=bulan ke-5, dan k_3=bulan ke-8) (k_1 = bulan ke-4, k_2 = bulan ke-6, k_3 = bulan ke-8, dan k_4 = bulan ke-11)

Pendekatan spline kubik dengan 1 knots belum optimal dan secara visual ditunjukkan pada Gambar 2. Pendekatan spline ini mengestimasi pola data yang *under estimate* pada bulan ke-10 dan pola data yang *over estimate* pada bulan ke-11. Hal yang sama juga tampak pada penggunaan 2 knots (Gambar 3) dan penggunaan 3 knots (Gambar 4). Berbeda dengan model estimasi spline kubik dengan 4 knots mampu mengestimasi pola data secara optimal dimana tampak dalam Gambar 6 tidak terjadi lagi *under estimate* pada bulan ke-10 dan pola data yang *over estimate* pada bulan ke-11, sehingga prilaku pola perubahan pada setiap interval bulan dapat teridentifikasi. Model spline dengan 4 titik knots optimal pada $k_1 = 4$, $k_2 = 6$, $k_3 = 8$, dan $k_4 = 11$ diberikan oleh:

$$\hat{y}(t) = 610,202 + 36,411t - 83,508t^2 + 17,486t^3 - 77,378(t-4)_+^3 + 121,429(t-6)_+^3 + \\ -92,575(t-8)_+^3 + 856,159(t-11)_+^3$$

Tabel 2. Nilai R^2 dan MSE dari Masing-masing Model Estimasi.

No	Model Estimasi	R^2	MSE
1	Regresi Parametrik Kubik	14,68 %	52.090,31
2	Regresi Nonparametrik Spline Kubik dengan k=6	61,19 %	23.692,50
3	Regresi Nonparametrik Spline Kubik dengan k_1=4 & k_2=5	63,64 %	22.200,14
4	Regresi Nonparametrik Spline Kubik dengan k_1=4, k_2=5 & k_3=8	64,07 %	21.933,36
5	Regresi Nonparametrik Spline Kubik dengan k_1=4, k_2=6, k_3=8 & k_4=11	68,64 %	19.141,67

Pada Tabel 2 disajikan nilai R^2 dan MSE untuk beberapa model pendekatan. Terlihat dari Tabel 2 bahwa model spline kubik dengan knots bulan ke-4, ke-6, ke-8, dan ke-11, mempunyai nilai R^2 yang paling besar yaitu 68,64 % dan MSE yang terkecil yaitu 19.141,67.

4. Kesimpulan

Kesimpulan yang dapat dituliskan dalam paper ini adalah bahwa pendekatan model spline *polynomial truncated* memiliki perhitungan matematika dan statistika yang mudah dan dapat menangani setiap perubahan pada pola data. Dan untuk mengestimasi model pada hubungan antara jumlah penjualan roti DEDE dengan waktu penjualan, dapat digunakan spline kubik *truncated* dengan empat knots optimal.

Daftar Pustaka

[1] Budiantara, I.N., 2001, Penentuan titik-titik knots dalam regresi spline, *Prosiding Seminar Nasional Matematika, FMIPA*, Universitas Negeri Yogyakarta.

[2] Budiantara, I.N., 2006, Model spline dengan knots optimal, *Jurnal Ilmu Dasar, FMIPA Universitas Jember*, 7, 77-85.

[3] Craven, P. dan Wahba, G., (1979), *Smoothing noise data with spline functions, Numerische Mathematics*, 31, 377-403.

[4] Eubank, R.L., 1988, *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

[5] Hardle, W., 1990, *Applied Nonparametric Regression*, Cambridge University Press, New York.

[6] Shao, J., 1993, Linear model selection by cross validation, *Journal of The American Statistical Association*, 88, 486 – 494.

[7] Wahba G., 1990, *Spline Models for Observational Data*, SIAM, Pensylvania.

[8] Wang, Y., 1998, Smoothing spline models with correlated errors, *Journal of The American Statistical Association*, 93, 343-348